SAMPLE DESIGN, ESTIMATES, AND RELIABILITY OF THE DATA

This section deals with design of the survey sample, weighting of responses, use of numerical factors to compensate for less than a full sample in making estimates, calculation of standard errors, and use of imputation flags.

Sample Design

The SIPP survey is based on a multi-stage stratified sample of the noninstitutional resident population of the United States. More specifically, the universe of the survey includes persons living in households, plus those persons living in group quarters such as college dormitories and rooming houses. In Wave 1 of the 1984 Panel, inmates of institutions, such as homes for the aged, and persons living abroad were not in the survey universe and thus not eligible for interview. Persons residing in military barracks, although part of the noninstitutional population, were also excluded from the survey universe in Wave 1. Other people in the Armed Forces were eligible, as long as they were living in a housing unit, whether off base or on.

For Wave 2 and subsequent waves, institutionalized persons, persons living abroad, and those living in military barracks become eligible for the survey only if they move into housing units in the United States with original sample persons, i.e., those who were interviewed in Wave 1.

Selection of Primary Sampling Units

To reduce sample selection and interviewing costs the Census Bureau first selects certain areas to be included in the sample, and then samples households within the selected areas. The first stage of this design involves the selection of these areas. The first step of this procedure is the definition of the United States in terms of counties or groups of counties called primary sampling units or PSU's.

PSU's with similar key socioeconomic characteristics are grouped together into strata. Then one sample PSU is selected from each stratum. The PSU's used for SIPP are a subsample of the sample PSU's used in the Current Population Survey.

Of the 174 strata into which PSU's were classified for the 1984 panel, 45 consisted of only a large single metropolitan area; these 45 areas were selected into the sample with certainty. These 45 PSU's are termed "self-representing." The remaining 129 strata consisted of 2 or more PSU's, from which only one was selected into the sample. These PSU's are termed "non-self-representing" because they were selected to represent other PSU's in their stratum as well as themselves.

The strata from which non-self-representing PSU's are selected typically cross State lines. For example, aside from the Detroit metro area, which represents itself, sampled PSU's in Michigan represent a geographically diverse area -- areas spread over the Midwestern States. (Thus, a tabulation of data coded to Michigan, for example, will not yield reasonable estimates for that State. Rather, State codes on the microdata files are primarily useful for determining applicable criteria for programs which vary from State to State.)

Selection of Ultimate Sampling Units

To arrive at the sample of households, geographic units called enumeration districts (ED's), with an average 350 housing units, are sampled from within each of these SIPP sample PSU's. Within those selected ED's 2 to 4 living quarters, or ultimate sampling units (USU's), are systematically selected from address lists prepared for the 1970 census. If the address lists are incomplete, small land areas are sampled. To account for living quarters built within each of the sample areas after the 1970 census, a sample is drawn of permits issued for construction of residential living quarters through March 1983. In jurisdictions that do not issue building permits, small land areas are sampled and the living quarters within are listed by field personnel and then subsampled. In addition, sample living quarters are selected from supplemental frames that include mobile home parks and new construction for which permits were issued prior to January 1, 1970, but for which construction was not completed until after April 1, 1970.

Sampling Rate and Weights

The objective of the sampling is to obtain a self-weighting probability sample. In a self-weighting sample, every sample unit has the same overall probability of selection. In self-representing PSU's the sampling rate is about 1 in 3,700. In non-self-representing PSU's, the sampling rate is higher, as the sampling is adjusted to account for the PSU's probability of selection. For example, if a non-self-representing PSU was selected with a probability of 1/10, the sampling rate within the PSU would be roughly 1 in 370 instead of 1 in 3,700.

In Wave 1, occupants of about 1,000 eligible living quarters were not interviewed because they refused to be interviewed, could not be found at home, were temporarily absent, or were otherwise unavailable. These households were not interviewed in Wave 2, and were classified as noninterviews because they were eligible for inclusion. Wave 2 included only 3 of the 4 rotation groups. For these reasons and as a result of following movers, a total of 14,532 living quarters were designated for Wave 2. Of these, 833 were not interviewed because they no longer contained eligible respondents. An additional 729 households were not interviewed in Wave 2 because of geographical remoteness or because of the reasons listed above for Wave 1 noninterviews. The noninterview rate for Wave 1 was 5 percent, and the combined noninterview rate for Wave 1 and Wave 2 was 9.4 percent.

estimation procedure used to derive SIPP person weights involves several stages of weight adjustments. In the first wave, each person received a base weight equal to the inverse of his/her probability of selection. In the second wave, each person received a base weight that accounted for differences in the probability of selection caused by the following of movers.

A noninterview adjustment factor was applied to the weight of each interviewed person to account for persons in occupied living quarters who were eligible for the sample but were not interviewed. A factor was applied to each interviewed person's weight to account for the SIPP sample areas not having the same population distribution as the strata from which they were selected.

An additional stage of adjustment to persons' weights was performed to bring the sample estimates into agreement with independent monthly estimates of the civilian (and some military) noninstitutional population of the United States by age, race, and sex. These independent estimates were based on statistics on births, deaths, immigration, and emigration; and statistics on the strength of the Armed Forces. To increase accuracy, weights were further adjusted in such a manner that SIPP sample estimates would closely agree with Current Population Survey (CPS) estimates by type of householder (married, single with relatives or single without relatives by sex and race) and relationship to householder (spouse or other). The estimation procedure for the data in the report also involved an adjustment so that the husband and wife of a household received the same weight.

The weight estimation procedure described above resulted in persons' weights varying from about 500 to 50,000. Persons in the sample for less than the entire 4-month period received zero weights for months not in the sample. Starting in Wave 5 the weighting system will also be adjusted to account for a reduction in the number of sample units interviewed. Most statistical software packages handle weighted data with no difficulty. In tabulating a characteristic the software takes each response and applies the person weight.

Figure 1 illustrates a simple example, in which 3 of 5 persons work full-time, 2 do not. But since the persons who do not work full-time happen to have higher weights than the others, weighted totals show the two groups to be equal.

FIGURE 1. Example of Weighted Data

			Raw Counts		Weighted Counts	
	Worked Full-Time	Weight	No	Yes	No	Yes
Person 1	No	4,000	1		4,000	
Person 2	No	5,000	1		5,000	
Person 3	Yes	3,000		1		3,000
Person 4	Yes	3,000		1		3,000
Person 5	Yes	3,000		1		3,000
			2	3	9,000	9,000

Preparing National Estimates for Persons, Families, and Households

Weights for persons are carried on each person record, on both the relational (hierarchical) and rectangular files. Weights for households and families are carried, respectively, on the household and family records of the relational file. The weighting process defines the weight of the household to be the same as the weight of the household reference person or househc.der, and the weight of a family or subfamily is that of the family or subfamily reference person.

On the rectangular file, where household, family, and subfamily segments appear on each person record, all of the applicable weights can be found in that record. Tallying household characteristics from every record would result in counting multi-person households more than once. One way to avoid estimating more households than there really are is to tally household characteristics using only the householder's record, since there is always one and only one householder per household. Similarly, the records of family or subfamily reference persons can be used in generating family and subfamily estimates.

Of course, many desired household characteristics are not already shown on household records or segments, but are derived by summarizing the characteristics of the persons in the household, as for example, the number of persons 65 years old and over in the household. Doing so with SIPP files is somewhat more complicated than with files in which person records are arranged in a strictly hierarchical fashion within household.

Household records in SIPP relational files carry pointers to each person who was a member of the household. There are five sets of pointers, one for each month of the reference period and one for the interview month. The rectangular file does not have these household-to-person pointers, but does identify the address ID of the household of which the person was a member each month. The file can be readily sorted on address ID within sample unit to group household members together for any particular reference month. Another option available to rectangular file users is to sort on the person number of the householder, provided on each household member's record.

Estimates for groups of persons other than households and families

Some analyses involve summarizing to units other than households or families. The persons within a household who benefit from food stamps are one such example. Only part of a family may receive aid or there may be two separate food stamp units living together. For each food stamp receiving unit one adult household member is designated as the prime recipient. The SIPP questionnaire also identifies which children and other household members are covered by those food stamps.

Food stamp coverage is recorded on the SIPP files in two ways. First, the primary recipient's record includes the person numbers of each household member covered, and each of the other covered persons' records has a flag that indicates membership in a food stamp receiving unit. Only the primary recipient's record specifies the amounts of food stamps received for the unit.

To tabulate the characteristics of all food stamp recipients in a household, the easiest approach might be to sort recipients together within households using the recipiency flags. But if it is necessary to discriminate between multiple food stamp receiving units within a household, the only way is to examine the primary recipient's record and use its list of person numbers to point to the secondary recipients in that unit. Then one could summarize appropriate characteristics across the person records. This way one could determine whether the food stamp recipiency unit includes a wage-earner, is part of a family below the poverty level, lives together with persons who are not covered, and so forth.

Other programs for which there are pointers from the primary recipient's record to other recipients in the household include Medicaid, AFDC, foster children payments, general assistance, health insurance, Railroad Retirement, Social Security and veterans payments. In all of these cases, all income received by the unit, including payments for the benefit of children, are reported on the record of the primary adult recipient and not on the records of secondary recipients. The weight of the primary recipient is most likely to be appropriate in tabulations of food stamp recipiency units and similar groups of individuals.

Estimates for Different Reference Periods

Each person and household is assigned 5 weights on each interview file, one for each of the four reference months and one for the interview month. Families and subfamilies are assigned only 4 weights since there is no attempt to define families as of the reference date. The 4 sets of reference month weights can be used only to form reference month estimates. Reference month estimates can be averaged, however, to form estimates of monthly averages over some period of time. For example, using the proper weights one can estimate the monthly average number of persons in a specified income range over the 4-month period.

The fifth weight is specific to the interview month. This weight can be used to form person or household estimates that specifically refer to characteristics as of the interview month. For example, one might want to estimate the number of unmarried adults living with an aged parent as of the latest observation. Interview weights can also be used to form estimates referring to the time period including the interview month and 4 previous months. One caution is that characteristics as of the interview date may not reflect that entire month--the persons could move, marry, or die before the end of the month.

The interview weight is also used for estimating a few of the demographic characteristics and other information that appear on the file for the 4-month reference period as a whole, but not for each month, such as race or sex.

None of these weights has been designed to yield the best estimates for a person's or household's status over two or more months, as for example, the number of households existing in October 1983 who experienced a 50 percent increase in income between July and August.

Calendar Month Data and Time Dimensioned Summary Statistics

In tabulating SIPP data for a particular calendar month, one must keep in mind the survey design. Most waves include 4 rotation groups, interviewed in four successive months. Figure 2 is a schematic diagram of the 1984 Panel design.

Months, quarters and years are shown along the top. Each cell shows the wave and rotation groups for which data are collected for each month. Thus, in the first interview, conducted in October 1983, data were collected from Wave 1-Rotation Group 1 households for the months of June, July, August and September.

As successive rotation groups are interviewed, the 4-month reference periods advance by 1 month. Wave 1-Rotation Group 2 households were interviewed in November 1983 for data for July through October.

In deriving calendar month or quarterly estimates from the data files, it is important to know how many rotation groups were interviewed, as less than the full sample may be available. If this is the case, the estimates must be inflated by an appropriate factor.

In some months, a full sample of 4 rotation groups from the same wave will be available. For Wave 1 (see figure 2), data for September 1983 were collected from the full sample. These data consist of month 4 data for Rotation Group 1, month 3 data for Rotation Group 2, month 2 data for Rotation Group 3, and month 1 data for Rotation Group 4. All of these figures (with appropriate weights) must be added together because any one rotation group includes only one-fourth of the SIPP sample.

In deriving quarterly estimates, a full sample consists of data for 4 rotation groups for each of the 3 months in the quarter. This would entail using data from 2 or 3 waves. For example, the fourth quarter of 1983 includes various rotation groups from Waves 1 and 2. Weighted data from all these rotation groups must be added together to form a full sample.

Note, however, that a full sample is not available for the third quarter of 1983. Or for subsequent quarters, the analyst may not want to wait for another wave of data to become available. Procedures to use in deriving estimates based on a partial sample are explained below.

Working With Less Than the Full Sample

Figure 2 also illustrates that for October 1983, data were collected from only three rotation groups of Wave 1. Thus the sample size available is three-fourths that available for September. The preferred way to handle this is to acquire Wave 2 as well, and combine October data for Wave 2-Rotation Group 1 with the Wave 1 October data for Rotation Groups 2, 3 and 4.

If a particular application does not require the full sample size, however, one could use only Wave 1 data for October and multiply weighted results by a factor of 1.33 to compensate for having only three-fourths of the sample. This is illustrated in figure 3.

FIGURE 3. Factors for Monthly Data: Wave 1, 1984 Panel

		Reference Period								
Month of	Rotation	Second Quarter			Third Quarter			Fourth Quarter		
Interview	Group	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
October	1			X	X	X	X			
November	2				X	X	X	X		
December	3					X	X	X	X	
January	4						X	X	X	X
Factors to Compensate for Missing Rotation Groups				4	2	1.33	1	1.33	2	4

To use Wave 1 data for the month of November, double the estimates (which compensates for having only one half of the sample consisting of Rotation Groups 3 and 4), and for December multiply the estimates by 4 (since they are based on a one-fourth sample consisting of rotation group 4 alone). Corresponding factors apply to data for June, July and August (also available in Wave 1) as well, and for these months the factors must be used, as the alternative of picking up the missing rotation groups in another wave does not exist.

A similar approach is applicable to subsequent waves as well. The particular factor to use is determined by the number of rotation groups covered in the time period one is analysing. Factors for Waves 1 and 2 and combined Wave 1 and 2 estimates are given in Table 1 below.

Table 1. Factors to be Applied to Basic Parameters to Obtain Parameters for Specific Reference Periods

Wave 1 Estimates

June 1983, December 1983	4.00
July 1983, November 1983	2.00
August 1983, October 1983	1.33
September 1983	1.00
3rd Quarter 1983	1.22
4th Quarter 1983	1.85
July-December 1983	1.06

Wave 2 Estimates

October 1983 and March 1984	4.00
November 1983 and February 1984	2.00
December 1983 and January 1984	1.33
4th Quarter 1983	1.85
1st Quarter 1984	1.85

Wave 1 and 2 Combined Estimates

June 1983 and March 1984	4.00
July 1983 and February 1984	2.00
August 1983 and January 1984	1.33
September through December 1983	1.00
3rd Quarter 1983	1.22
4th Quarter 1983	1.00
1st Quarter 1984	1.85
July-December 1983	1.06

Factors must also be applied to quarterly estimates or those for longer periods of time if less than the full sample for any month is available. Thus, in table 1 a factor of 1.22 must be applied to third quarter 1983 estimates, 1.85 to fourth quarter estimates using either Wave 1 or Wave 2, but a factor of 1.00 (i.e., no factor is needed) for fourth quarter 1983 estimates using full sample data from the combined Wave 1 and Wave 2 files.

Caveats for Calendar Month Data

Although it is possible to examine the data on a monthly basis and examine the data in a strictly cross sectional sense, there are qualifications or biases in this type of analysis.

First, no evaluations have been made of responses to income and related variables that are provided on a monthly basis. There may be some biases in this reporting. For example, people may tend to report a rough monthly average for their income over the four month reference period rather than specifically recalling amounts separately for each month. If this were so it would not be possible to analyze real month-to-month changes in income figures.

Second, most data users have been able to work only with annual income figures to this point, using the census, CPS or other surveys which measure income only once during a year. There will be considerable temptation for SIPP users to return to familiar analytical ground by multiplying 4-month income figures by 3 to estimate 12-month income. To do so would ignore seasonal variation in employment and income. A better approach to annual income would be to match together the first several waves and look at actual income experience across 12 months, perhaps comparing the results to the annual income and taxation information reported in Wave 5.

Time-Dimensioned Summary Statistics

An approach to analyzing these data that would reduce the biases just discussed for monthly estimates involves summarizing data across time. In this approach one calculates standard summary statistics such as counts, means, and modes across time as well as across individuals.

For example, instead of calculating the number of persons with incomes over $3,000 for the month of July, one would calculate the number of persons with a mean monthly income of $3,000 or more during the 3rd quarter.

This approach is relatively straightforward at the person level. However, at the family or household level, an additional complexity is added. One must first define these groups and identify the changes that occur during the quarter.[1] Then the conditions under which new groups are formed must be defined. Longitudinal concepts of households and families are the subject of a Working Paper, "Toward a Longitudinal Definition of Households" by David McMillen and Roger Herriot, available from the Census Bureau.

Producing Estimates Below the National Level

Census Regions

The total estimate for a region is the sum of the state estimates in that region. However, one of the groups of states, formed for confidentiality reasons, crosses regional boundaries. This group consists of South Dakota

[1] These problems do not arise in Wave 1, as households were defined as of the interview and changes during the reference months were not recorded.

(Midwest Region), Idaho (West Region), New Mexico (West Region), and Wyoming (West Region). To compute the total estimate for the Midwest Region, a factor of .203 should be applied to the above group's total estimate and added to the sum of the other state estimates in the Midwest Region. For the West Region, a factor of .797 should be applied to the above group's total estimate and added to the sum of the other states in the West.

Estimates for regions included in the published SIPP reports reflect the actual region of residence, not the results of proration across the 4-state group. Thus there will be minor discrepancies between published regional totals and estimates derivable from microdata files for the Midwest and West regions.

Estimates from this sample for individual states are subject to very high variance and are not recommended. The State codes on the file are primarily of use for linking respondent characteristics with appropriate contextual variables (e.g., State-specific welfare criteria) and for tabulating data by user-defined groupings of States.

Producing Estimates for the Metropolitan Population

For 15 states in the SIPP sample, metropolitan or nonmetropolitan residence is identified. (On the rectangular file, use variable H*-METRO, characters 94, 382, 670, and 958. On the relational file, use METRO, character 24 on the household record). Metropolitan residence is defined according to the definition of Metropolitan Statistical Areas as of June 30, 1983. In 21 additional states, where the nonmetropolitan population in the sample was small enough to present a disclosure risk, a fraction of the metropolitan sample was recoded so as to be indistinguishable from nonmetropolitan cases (METRO=2). In these states, therefore, the cases coded as metropolitan (METRO=1) represent only a subsample of that population.

In producing state estimates for a metropolitan characteristic, multiply the individual, family, or household weights by the metropolitan inflation factor for that state, presented in Table 2 below. (This inflation factor compensates for the subsampling of the metropolitan population and is 1.0 for the states with complete identification of the metropolitan population).

In producing regional or national estimates of the metropolitan population it is also necessary to compensate for the fact that no metropolitan subsample is identified within two states (Maine and Iowa) and one state-group (Mississippi-West Virginia). (There were no metropolitan areas sampled in South Dakota-Idaho-New Mexico-Wyoming). Therefore, a different factor for regional and national estimates is in the right-hand column of Table 2 below. The results of regional and national tabulations of the metropolitan population will be biased slightly, although less than one-half of one percent of the metropolitan population is not represented.

Table 2. Metropolitan Subsample Factors

(Multiply these factors times the weight for the person,
family or household)

		Factors for use in State or MSA Tabulations	Factors for use in Regional or National Tabs
Northeast:	Connecticut	1.0390	1.0432
	Maine	-	-
	Massachusetts	1.0000	1.0040
	New Jersey	1.0000	1.0040
	New York	1.0110	1.0150
	Pennsylvania	1.0025	1.0065
	Rhode Island	1.2549	1.2599
Midwest:	Illinois	1.0232	1.0310
	Indiana	1.0000	1.0076
	Iowa	-	-
	Kansas	1.6024	1.6146
	Michigan	1.0000	1.0076
	Minnesota	1.0000	1.0076
	Missouri	1.0611	1.0692
	Nebraska	1.7454	1.7587
	Ohio	1.0134	1.0211
	Wisconsin	1.0700	1.0782
South:	Alabama	1.1441	1.1511
	Arkansas	1.0000	1.0061
	Delaware	1.0000	1.0061
	District of Columbia	1.0000	1.0061
	Florida	1.0333	1.0396
	Georgia	1.0000	1.0061
	Kentucky	1.1124	1.1192
	Louisiana	1.1470	1.1540
	Maryland	1.0000	1.0061
	North Carolina	1.0000	1.0061
	Oklahoma	1.1146	1.1214
	South Carolina	1.1270	1.1339
	Tennessee	1.0000	1.0061
	Texas	1.0192	1.0254
	Virginia	1.0778	1.0844
	West Va. - Miss.	-	-
West:	Arizona	1.0870	1.0870
	California	1.0000	1.0000
	Colorado	1.0000	1.0000
	Hawaii	1.0000	1.0000
	Oregon	1.0879	1.0879
	Washington	1.0868	1.0868

- indicates no metropolitan subsample is shown for the State.

Estimates for the metropolitan population produced from the microdata files will differ from published summary figures for the metropolitan population not only because of the subsampling scheme but also because of differences in the definition of the metropolitan population. Published figures are based on Standard Metropolitan Statistical Areas (SMSA's) defined as of June 30, 1981, consistent with the definition for the 1980 census. The microdata files use the definitions for Metropolitan Statistical Areas(MSA's) as of June 30, 1983. That definitional change resulted in increasing the metropolitan population by 1.4 percent. Eventually, the published figures will also reflect 1983 MSA definitions.

Producing Estimates for the Nonmetropolitan Population

State, regional, and national estimates of the nonmetropolitan population cannot be computed directly, except for the 15 states where the factor in Table 2 is 1.0. In all other states, the cases identified as not in the metropolitan subsample (METRO=2) are a mixture of nonmetropolitan and metropolitan households. Only an indirect method of estimation is available: first compute an estimate for the total population, then subtract the estimate for the metropolitan population.

Codes for Individual MSA's

Codes for certain large individual MSA's are included on the microdata files, much as are State codes, to provide users some flexibility in defining higher level aggregate areas and to allow linking respondent characteristics to available contextual variables. Individual MSA codes are given if the MSA has at least 250,000 inhabitants in sampled counties within the state, and if its identification would not result in the indirect identification of residual metropolitan population less than 250,000. Sample sizes associated with individual MSA's are typically very small.

When creating estimates for particular identified MSA's or CMSA's apply the Table 2 factor to the weights appropriate to the state, as discussed above. For multi-state MSA's, use the factor appropriate to each state part. For example, to tabulate data for the Washington, DC-MD-VA MSA, apply the Virginia factor of 1.0778 to weights for residents of the Virginia part of the MSA; Maryland and DC residents require no modification to the weights (i.e., their factors equal 1.0). This may still not produce reasonable estimates for an individual MSA for three reasons: 1) the sample size is very small; 2) the MSA may be non-self-representing; and 3) certain counties added to MSA's between 1970 and 1983 may not have been included in the 1984 panel.

Sampling Variability

Data found in SIPP publications or in user tabulations from the SIPP microdata are estimates based on the weighted counts from the survey. These numbers only approximate the far more costly counts that would result from a census of the entire population from which the sample was drawn. There are two types of errors possible in an estimate based on a sample survey: Sampling and nonsampling. We are able to provide estimates of the magnitude of SIPP sampling error, but this is not true of nonsampling error.

Standard Errors and Confidence Intervals

Standard errors indicate the magnitude of the sampling error. They also partially measure the effect of some nonsampling errors in response and enumeration, but do not measure any systematic biases in the data. The standard errors for the most part measure the variations that occurred by chance because a sample was surveyed instead of the entire population.

The sample estimate and its standard error enable one to construct confidence intervals, ranges that would include the average result of all possible samples with a known probability. For example, if all possible samples were selected, each of these being surveyed under essentially the same general conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then:

1. Approximately 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the average result of all possible samples.

2. Approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average result of all possible samples.

3. Approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with a specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

Hypothesis Testing

Standard errors may also be used for hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common types of hypotheses tested are 1) the population parameters are identical versus 2) they are different. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the parameters are different when, in fact, they are identical.

To perform the most common test, let X_A and X_B be sample estimates of two parameters of interest. A subsequent section explains how to derive a standard error on the difference X_A-X_B. Let that standard error be S_{DIFF}. Compute the ratio $R=(X_A-X_B)/S_{DIFF}$. If this ratio is between -2 and +2, no conclusion about the parameters is justified at the 5 percent significance level. If on the other hand, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 5 percent level.

In this event, it is a commonly accepted practice to say that the parameters are different. Of course, sometimes this conclusion will be wrong. When the parameters are, in fact, the same, there is a 5 percent chance of concluding that they are different.

Calculating Standard Errors for SIPP

There are two ways for users to compute a standard error for SIPP estimates. One method is to compute variances directly using half-sample and pseudostratum *codes. A second method involves calculating generalized standard errors using simple charts and formulas found in published reports or microdata documentation.

Generalized Standard Errors

To derive standard errors that are applicable to a wide variety of statistics and can be prepared at a moderate cost, a number of approximations are required. Most of the SIPP statistics have greater variance than those obtained through a simple random sample because clusters of living quarters are sampled for SIPP.

Two parameters, denoted "a" and "b", have been developed to calculate variances for each type of characteristic. These "a" and "b" parameters, found in table 3, are used in estimating standard errors of survey estimates, and these standard errors are referred to as generalized standard errors.

All statistics do not have the same variance behavior; "a" and "b" parameters were computed for groups of statistics with similar variance behavior. The parameters were computed directly from SIPP 3rd quarter 1983 data

*Revised May 1985

Table 3. SIPP 1984 Generalized Variance Parameters

Characteristic	Basic Parameters	
	a	b
16+ total persons: program participation and benefits	-0.00009428	16059
As above for 16+ total males	-0.00019844	16059
As above for 16+ total females	-0.00017961	16059
16+ total persons: income, labor force	-0.00003214	5475
As above for 16+ total males	-0.00006765	5475
As above for 16+ total females	-0.00006123	5475
0+ Total persons: all items	-0.00008637	19911
As above for total males	-0.00017863	19911
As above for total females	-0.00016724	19911
Black persons: all items	-0.00026695	7366
As above for Black males	-0.00057368	7366
As above for Black females	-0.00049929	7366
Total households: all items	-0.00007644	6766
Black households: all items	-0.00046611	4675

The "a" and "b" parameters may be used to approximate the standard error for estimated numbers and percentages. Because the actual increase in variahce was not identical for all statistics within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error rather than the precise standard error for any specific statistic. That is why we refer to these as generalized standard errors.

Computing Variances Directly

Psuedo half-sample codes and psuedostratum codes (assigned in such a way as to avoid any disclosure risk) are included on the file to enable direct computation of variances by methods such as balanced repeated replications.[2] This method may be used if the user can not use the generalized standard errors, as in computing the variance of a correlation coefficient between, say, interest income and dividend income. Since a number of statistical software packages provide simple procedures for using half-sample codes, you may consult documentation for your statistical software for further discussion. The Census Bureau, however, does not vouch for the appropriateness or accuracy of such software.

Variances computed directly will vary from variances estimated by the Census Bureau. These differences are a result of the use of artificial stratum codes on the public use file, whereas the Census Bureau has access to the actual stratum identifiers. Actual stratum codes are withheld from the public-use microdata so as to avoid identifying geographic areas so small that they risk disclosure of confidential information.

Even though these are artificial stratum codes, the variance estimates are expected to be similar to those produced by the Bureau using the real stratum codes. This method is involved, may be expensive, and, of course, is available only to users of SIPP microdata, not users of SIPP publications.

Examples Using Generalized Standard Errors

Some examples illustrate the use of "a" and "b" parameters in Table 3 for computing a standard error and the corresponding confidence intervals.

Standard Error of Total

The formula for computing the standard error for a total is:

$$s = \sqrt{ax^2 + bx} \qquad (1)$$

[2] William G. Cochran provides a list of references discussing the application of this technique in Sampling Techniques, 3rd Ed. (New York: John Wiley and Sons, 1977), p. 321.

.ere "a" and "b" are the parameters associated with the estimate for the particular reference period and x is the weighted estimate.

Based on a tabulation from the SIPP survey data you would find that there were 16,000,000 households with a mean monthly income during the 3rd quarter of 1983 of $3,000 and over. Suppose you want to develop a 95% confidence interval so you can assess how precise the estimate of 16,000,000 is.

Step 1:

Determine the appropriate "a" and "b" parameters by looking them up in table 3. Since we are dealing with income data for all households the "a" and "b" parameters are -.00007644 and 6766.

Step 2:

Enter these figures in the above formula

$$s = \sqrt{ax^2 + bx}$$

$$= \sqrt{(-.00007644) \times (16{,}000{,}000)^2 + (6766 \times 16{,}000{,}000)}$$

$$= 297{,}804.231$$

where 16,000,000 is the estimate, and -.00007644 and 6766 are the "a" and "b" parameters. The resulting standard error (rounded off) is 297,804.

Step 3:

To determine the 95% confidence interval of the estimate, multiply 2 times the standard error, yielding 595,608. The lower bound of the confidence interval is then 16,000,000 minus 595,608 or roughly 15.4 million, and the upper bound is 16,000,000 plus 595,608 or roughly 16.6 million.

Thus we can conclude with 95% confidence that the average estimate derived from all possible samples lies within the interval of 15.4 million to 16.6 million.

The foregoing example assumes you are working with the full SIPP sample, as will normally be the case with SIPP reports and user tabulations. But if you are making a tabulation from SIPP microdata for a reference month for which you do not have data for all rotation groups, you must weight the estimate up by an appropriate factor to compensate for the smaller sample size; you must similarly adjust the estimates of variance.

When you are working with fewer than all 4 rotation groups, the formula becomes

$$s = \sqrt{ax^2 + bx} \cdot \sqrt{f} \qquad (2)$$

where the first part of the expression is the same as before, and "f" is a factor compensating for sample size. In other words, when the estimate is weighted up by a factor, the standard error must be multiplied by the square root of the same factor.

The "f" factors for various reference periods are found in table 1 above. The standard error in the above example was 297,804. If we were working with data for July 1983, a month covered by only the first two rotation groups in Wave 1 (see figure 2), our initial estimate using the weights on the microdata file might have been 8,000,000. To compensate for the 2 missing rotation groups, we would apply the factor of 2.0, and thereby double our estimate to 16,000,000. The same factor would enter into the formula in equation (2) to give

$$s = 297{,}804 \times \sqrt{2.0} = 421{,}158$$

as the standard error of an estimated 16,000,000 based on 2 rotation groups instead of 4. The confidence interval is then determined in the same way, using this revised standard error.

Wave 1 represents a special case because there are 3 reference months at the start of the survey when the survey did not yet cover all four rotation groups. Only one rotation group has data for June 1983, two for July 1983, and three for August 1983. The first SIPP report included data for the third quarter 1983.

For that period of partial coverage a factor of 1.22 is appropriate, as shown in table 1. If wave 1 data were used to estimate the 4th quarter, the factor would be 1.85. Of course, wave 2 supplies the missing rotation groups for that quarter. If wave 1 and wave 2 files were used together, estimates could be made from the full sample, so that no factor adjustment would be needed. Since the factors associated with the metropolitan area subsample are generally very close to 1.0, the factors may be ignored in calculating variances for metropolitan summaries.

Standard Error of a Percent

Computing the standard error and confidence interval for a percent follows a similar procedure. The formula for the generalized standard error of a percent is:

$$s = \sqrt{\frac{b}{y}\, p(100-p)} \cdot \sqrt{f} \qquad (3)$$

where

y = the base of the percent (use weighted estimate), i.e., the size of the subclass of interest,

p = the percentage of persons, families, or households possessing the characteristic of interest, b = the larger of the "b" parameters for the numerator and denominator, and,

f = the factor to adjust for missing rotation groups if necessary.

Note that the "a" parameter is not used.

Suppose we find that of the households in Wave 1 who had a mean monthly income of $3,000 and over in the third quarter of 1983, 8,916,000 (8.6%) were black. To construct a 95% confidence interval, follow the steps shown below.

Step 1:

Examine the "b" parameter in table 3 for both total and black households to determine the larger of the two. In this case the "b" parameter for total households, 6766, is larger.

The "f" factor from table 1 that is applied to the base parameters to adjust for incomplete data is 1.22, applicable to 3rd quarter data.

Step 2:

Entering the values into the formula in equation (3):

$$s = \sqrt{\frac{6766}{8,916,000}(8.6)(100-8.6)} \cdot \sqrt{1.22}$$

provides us with a standard error of 0.85 percent.

Step 3:

Multiplying the standard error by 2 and adding and subtracting this quantity from the estimate of 8.6% provides a 95% confidence interval of 6.9% to 10.3%.

Standard Error of a Difference

The standard error of a difference between two sample estimates is approximately equal to

$$s_{(x-y)} = \sqrt{s_x^2 + s_y^2 - 2rs_xs_y} \qquad (4)$$

where s_x and s_y are the standard errors of the estimates x and y. The estimates can be numbers, percents, ratios, etc. The correlation between x and y is denoted by the correlation coefficient r.[3] Table 4 presents the correlation coefficients r for comparisons between months and between quarters. For other types of comparisons, assume r equals zero if it is believed that the value of one variable does not give a strong indication of the value of the other variable. If r is really positive then this assumption will lead to overestimates of the true standard error. If r is negative, the result will be an underestimate of the actual standard error.

As an illustration, SIPP estimates show that the number of persons in nonfarm households with mean monthly household cash income over $4,000 during the third quarter of 1983 who were aged 35-44 years was 5,313,000 and the number of those aged 25-34 years was 4,353,000, an estimated difference of 960,000. Using the Wave 1 parameters a=-.00003214, b=5475, and f=1.22 in equation (2), the standard errors of the estimates for each age group are 185,422 and 168,324 respectively. It is reasonable to assume that these two estimates are not highly correlated. Therefore, the standard error of the estimated difference of 960,000 is

$$\sqrt{(185,422)^2 + (168,324)^2} = 250,428$$

Suppose that it is desired to test the estimated difference at the 95 percent confidence level. The estimated difference divided by the standard error of the difference, 960,000/250,428, is 3.83. Since this is greater than 2 it is concluded that the difference is significant at the 95 percent confidence level.

Standard Error of a Mean

A mean is defined here to be the average quantity of some item (other than persons, families, or households) per person, family, or household. For example, it could be the average monthly household income of females aged 25 to 34. The standard error of a mean can be approximated by the formula below. Because of the approximations used in developing the formula, an estimate of the standard error of the mean obtained from that formula will generally underestimate the true standard error. The formula used to estimate the standard error of a mean $\bar{x}$ is

$$s_{\bar{x}} = \sqrt{\frac{b}{y} s^2} \cdot \sqrt{f} \qquad (5)$$

[3] The correlation coefficient measures the extent to which the value of one variable gives an indication of the value of another variable. An example of a positive correlation is that between food stamp and AFDC recipiency. Food stamp and bond income recipiency are variables possessing a negative correlation. Another example of variables with positive correlation occurs when it is desired to measure the difference in a variable between two months or quarters.

Table 4. Correlations for Monthly and Quarterly Comparisons

Wave 1 Estimates	Total income, wage income and similar types of income	Program participation income, nonincome, labor force
Jun-Jul, Nov-Dec 1983	0.57	0.35
Jul-Aug, Oct-Nov 1983	0.65	0.41
Aug-Sep, Sep-Oct 1983	0.69	0.43
Jun-Aug, Oct-Dec 1983	0.43	0.26
Jul-Sep, Sep-Nov 1983	0.53	0.32
Aug-Oct 1983	0.50	0.30
Jun-Sep, Sep-Dec 1983	0.35	0.20
Jul-Oct, Aug-Nov 1983	0.29	0.16
Jun-Oct, Jul-Nov, Aug-Dec, Jun-Nov, Jul-Dec, Jun-Dec 1983	0.00	0.00
3rd Quarter-4th Quarter 1983	0.28	0.14
Wave 2 Estimates		
Oct-Nov 1983, Feb-Mar 1984	0.57	0.35
Nov-Dec 1983, Jan-Feb 1984	0.65	0.41
Dec 1983-Jan 1984	0.80	0.50
Oct-Dec 1983, Jan-Mar 1984	0.43	0.26
Nov 1983-Jan 1984, Dec 1983-Feb 1984	0.61	0.37
Oct 1983-Jan 1984, Dec 1983-Mar 1984	0.40	0.23
Nov 1983-Feb 1984	0.35	0.20
Oct 1983-Feb 1984, Nov 1983-Mar 1984 Oct 1983-Mar 1984	0.00	0.00
4th Quarter 1983-1st Quarter 1984	0.34	0.20

Table 4--Continued

Wave 1 and 2 Combined Estimates	Total income, wage income and similar types of income	Program participation income, nonincome, labor force
Jun-Jul 1983, Feb-Mar 1984	0.57	0.35
Jul-Aug 1983, Jan-Feb 1984	0.65	0.41
Aug-Sep 1983, Dec 1983-Jan 1984	0.69	0.43
Sep-Oct, Oct-Nov,Nov-Dec 1983	0.80	0.50
Jun-Aug 1983, Jan-Mar 1984	0.43	0.26
Jul-Sep 1983, Dec 1983-Feb 1984	0.53	0.32
Aug-Oct 1983, Nov 1983-Jan 1984	0.65	0.39
Sep-Nov, Oct-Dec 1983	0.75	0.45
Jun-Sep 1983, Dec 1983-Mar 1984	0.35	0.20
Jul-Oct 1983, Nov 1983-Feb 1984	0.50	0.28
Aug-Nov 1983, Oct 1983-Jan 1984	0.61	0.35
Sep-Dec 1983	0.70	0.40
Jun-Oct 1983, Nov 1983-Mar 1984	0.33	0.18
Jul-Nov 1983, Oct 1983-Feb 1984	0.46	0.25
Aug-Dec 1983, Sep 1983-Jan 1984	0.56	0.30
Jun-Nov 1983, Oct 1983-Mar 1984	0.30	0.15
Jul-Dec 1983, Sep 1983-Feb 1984	0.42	0.21
Aug 1983-Jan 1984	0.60	0.30
Jun-Dec 1983, Sep 1983-Mar 1984	0.28	0.13
Jul 1983-Jan 1984, Aug 1983-Feb 1984	0.45	0.20
Jun 1983-Jan 1984, Aug 1983-Mar 1984	0.29	0.12
Jul 1983-Feb 1984	0.25	0.10
Jun 1983-Feb 1984, Jul 1983-Mar 1984	0.00	0.00
Jun 1983-Mar 1984		
3rd Qcarter-4th Quarter 1983	0.63	0.36
4th Quarter 1983-1st Quarter 1984	0.51	0.29
3rd Quarter 1983-1st Quarter 1984	0.39	0.18

where y is the size of the base, s^2 is the estimated variance of x, b′ is the parameter associated with the particular type of item, and f is the adjustment factor.

The estimated population variance, s^2, is given by formula (6):

$$s^2 = \frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i} - \bar{x}^2 \qquad (6)$$

where there are n persons with the item of interest; w_i is the final weight for person i; and x_i is the value of the estimate for person i.

If the calculation of s^2 using formula (6) is too cumbersome, then formula (7) may be used instead:

$$s^2 = \sum_{i=1}^{c} p_i x_i^2 - \bar{x}^2 \qquad (7)$$

where each person (or other unit of analysis) is in one of c groups (e.g., income categories within an income distribution); the p_i's are the estimated proportions of responses within each group; the x_i's are the midpoints of each group. If group c is open-ended, i.e., no upper interval boundary exists, then an approximate average value is

$$x_c = \frac{3}{2} Z_{c-1} \qquad (8)$$

where Z_{c-1} is the lower boundary of the group (e.g., $75,000 in the category $75,000 or more). If an open-ended group c does exist, the approximation could easily be bad. To reduce this danger, create data categories so as to keep c and Z_{c-1} large. This could be done by creating more categories, e.g., more income groups.

Standard Error of a Mean Number of Persons with Characteristic Per Family or Household

Mean values for persons in families or households may be calculated as the ratio of two numbers. The denominator, y, represents a count of families or households of a certain class, and the numerator, x, represents a count of persons with the characteristic under consideration who are members of these families or households. For example, the mean number of children per family with children is calculated as

$$\frac{x}{y} = \frac{\text{total number of children in families}}{\text{total number of families with children}}$$

For means of this kind, the standard error is approximated by the following formula:

$$s_{\left(\frac{x}{y}\right)} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{s_y}{y}\right)^2 + \left(\frac{s_x}{x}\right)^2 - 2r\left(\frac{s_x}{x}\right)\left(\frac{s_y}{y}\right)\right]} \qquad (9)$$

The standard error of the estimated number of families or households is s_y, and the standard error of the estimated number of persons with the characteristic is s_x. In the formula, r represents the correlation coefficient between the numerator and the denominator of the estimate. If at least one member of each family or household in the class possesses the characteristic of interest, then use 0.7 as an estimate of r. If, on the other hand, it is possible that no member of a family or household has the characteristic, then use r = 0. In the example, you would use r = 0.7 for the average number of persons per family, but r = 0 for the average number of teenagers per family.

Standard Error of a Median

To compute a median, first group the units of interest (e.g., persons) into cells by the value of the statistic under consideration (e.g., single years of age). Then form a cumulative density for the cells (e.g., by cumulatively adding the proportion of persons of each age). Identify the first cell with cumulative density greater than 0.5. Use interpolation to find the value of the characteristic that corresponds to cumulative density 0.5. That value is the estimated median. Different methods of interpolation may be used. The most common are simple linear interpolation and pareto interpolation. No universal rules exist on which method to use. The best procedure is to define the cells (e.g., income intervals) to be so small that the method of interpolation does not matter.

The sampling variability of an estimated median depends upon the form of the distribution as well as the size of its base or class. Given that the data were grouped into intervals (e.g., income intervals), then the standard error of a median is given by

$$\frac{\sqrt{bN}\,(A_2 - A_1)}{2(N_2 - N_1)} = \frac{\sqrt{bN}\,W}{2F} \qquad (10)$$

or

$$\frac{\sqrt{b}\,N \ln(A_2/A_1)}{\sqrt{N}\,\ln\left[(N-N_1)/(N-N_2)\right]} \qquad (11)$$

depending on whether the linear (10) or the Pareto (11) interpolation was used for estimating the median, where

M = the estimated median

A_1 and A_2 = the lower and upper boundaries of the interval in which the median falls,

W = $A_2 - A_1$, the width of the interval in which the median falls,

N_1 and N_2 = the number of units with the characteristic (e.g., income) less than A_1 and A_2, respectively,

F = $N_2 - N_1$, the number of units in the interval in which the median lies,

N = the total number of units in the frequency distribution,

b = the appropriate value of the parameter "b".

The following example illustrates the computation of the standard error of a median using linear interpolation. SIPP estimates from the report, "Economic Characteristics of Households in the United States: Third Quarter 1983," Series P-70, No. 1, table 1, show that the estimated median of the average monthly household cash income of females in the third quarter of 1983 was $1,841 and N = 115,848,000. The appropriate "b" parameter from table 3 of this chapter is 19,911, which must be multiplied by the 3rd quarter factor of 1.22, yielding 24,291. We used the interval defined by A_1 = $1,600, A_2 = $1,999, N_1 = 50,084,000, and N_2 = 62,087,000. So W = $399 and F = 12,003,000. Using the formula in equation (10) above the approximate standard error is

$$\sqrt{\frac{(24{,}291)\ (115{,}848{,}000)\ (\$399)}{2\ (12{,}003{,}000)}} = \$27.88 \qquad (12)$$

Thus, rounding to $28, the 68 percent confidence interval of the median is from $1,813 to $1,869, and the 95 percent confidence interval is from $1,785 to $1,897.[4]

[4] The standard error of $27.88 computed here differs from the standard error of the median found in the report referenced in the text. Since publication of the report, new parameters in table 3 of this chapter were developed based entirely on SIPP data. These parameters, given in this chapter, are to be used in place of those given in the Source and Reliability sections of that report or the Wave 1 Technical Documentation.

Standard Errors of Ratios of Means or Medians

In this section, the correlation between the numerator and denominator, r, is assumed to be zero. So, the standard error for a ratio of means or medians is approximated by this formula:

$$s_{\left(\frac{x}{y}\right)} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{s_y}{y}\right)^2 + \left(\frac{s_x}{x}\right)^2\right]} \qquad (13)$$

The standard errors of the two means or medians are s_x and s_y. If r is actually positive (negative), then this procedure will provide an overestimate (underestimate) of the standard error for the ratio of means and medians.

Nonsampling Error

In addition to sampling error, discussed above, nonsampling errors are also present in SIPP data. Nonsampling errors can be attributed to many sources.

Undercoverage

Some housing units may have been missed in the listing operation prior to sampling; sometimes persons are missed within a sampled household. Past studies of censuses and household surveys have shown that undercoverage varies by age, race, and residence. Ratio estimation to independent age-sex-race population controls partially corrects for the bias due to survey undercoverage. However, biases exist in those estimates insofar as the characteristics of missed persons differ from those of respondents in each age-sex-race group. Further, the independent population controls have not been adjusted for undercoverage in the decennial census. Undercoverage in SIPP relative to the independent controls is about 7 percent for both Wave 1 and Wave 2. The undercoverage rate is likely to increase in subsequent waves due to lack of complete coverage of immigrants, institutional discharges, and movers from military barracks.

Respondent and Enumerator Error

Persons may have misinterpreted certain questions, or there may be an inability or unwillingness to provide the correct information. One source of such inability arises when one household member responds for other members. In another, a number of evaluation programs from the decennial census have suggested that some persons tend to underreport their income. Or, there may be a problem in recalling information, though the shorter reference period employed in SIPP should reduce this problem. The greater detail in SIPP questions and the training of interviewers should help prompt more complete income reporting than in other surveys.

Processing Error

Errors may have been introduced in the handling of the questionnaires by the Census Bureau. The coding of write-in entries for occupation, for instance, is subject to a certain level of mistakes.

Nonresponse

Nonresponse to particular questions in the survey also allow for the introduction of bias into the data, since the characteristics of nonrespondents may differ from those of respondents.

The initial evaluation of the quality of the data from SIPP show improvements in the accuracy and completeness of the data on income and program participation over that obtained from March CPS. For the third quarter of 1983, SIPP nonresponse rates ranged from a low of about 3 percent for questions about Aid to Families with Dependent Children and food stamp allotments, to about 13 percent for those concerning self-employment income. These rates contrast sharply with the higher nonresponse rates from the March CPS. The rates for CPS range from a low of 9 percent for food stamp allotments to 24 percent for self-employment income.

The reasons attributed to the improvement in the measurement of income are SIPP's shorter recall period, and more emphasis in SIPP on complete and accurate reporting of income data. For example, in determining assets respondents are asked about type of ownership (whether jointly held) as well as value. Respondents are called back when information is incomplete.

The nonresponse rate for monthly wage and salary income overall averaged about 6.2 percent for the initial SIPP interviews. However, proxy responses caused significantly higher nonresponse rates for some of the key items.

The nonresponse rate for self-respondents, which accounted for 64 percent of the total, was 4.6 percent, while the rate for proxy respondents was 9.0 percent.

Noninterview rates for the first two waves of SIPP are 4.8 percent for Wave 1 and 9.4 percent for Wave 1 and Wave 2 combined. Most of these cases (77 percent) were refusals, but other cases included "no one at home" and "temporarily absent". These rates are an improvement on the rates experienced in the Income Survey Development Program (ISDP), a predecessor to SIPP, and are comparable with rates obtained in CPS. Since SIPP does not replenish a panel in the same manner as CPS, the SIPP noninterview rate will climb considerably above the monthly CPS rate. The Bureau has used complex techniques to adjust the weights for nonresponse, but the success of these techniques in avoiding bias is unknown.

Data quality issues in SIPP are also discussed in "Economic Characteristics of Households in the United States: Fourth Quarter 1983," Series P-70-83-4, Appendix D. This appendix includes comparisons of nonresponse in SIPP and the

March 1984 CPS, as well as comparisons of estimates derived from SIPP with independent estimates for several income types.

Imputation

There are almost no missing data on SIPP microdata files. Nonresponse by an entire household is dealt with in the weighting procedures. That is, noninterviewed households are given zero weights and interviewed households are weighted up to compensate. When an individual within the household refuses the interview or when a response to an individual question is missing, beginning with Wave 2, census computers make imputations for the missing data. For Wave 1, nonresponse to an entire questionnaire by an individual caused the household to receive a zero weight. If the person answered a certain minimum group of questions in Wave 1, the responses to the other items were imputed. Imputations involve the replacement of missing data after Wave 1 with a corresponding value from a housing unit or person having certain other characteristics in common with the unit or person in question.

In general this imputation procedure enhances the usefulness of the data. It simplifies processing for the microdata user by eliminating "not reported" categories. Imputation also enhances the accuracy of the data on targeted characteristics. By imputing a missing characteristic with that of someone similar in other key aspects, the user can work with a more complete data set. When an imputed characteristic is aggregated over a sizable number of persons, deviations from actual (unknown) values tend to even out. Using imputed values also yields more accuracy than substituting the mean for missing data, since the mean would be based on persons perhaps substantially different from those with the missing items. On the other hand, use of imputed values can harm the accuracy of characteristics that were not targeted. The targeted characteristics concern socioeconomic stratum.

Inclusion of Imputation Flags

If the characteristics of nonrespondents are systematically different from the characteristics of respondents, as may well be the case for income variables, then it is possible that the imputation system masks certain biases due to nonresponse. For this reason the SIPP microdata files include flags for many data items which allow the user to discriminate between those responses which were actually reported and those entries which were supplied through imputations. These flags, or imputation indicators, appear at the end of the house hold, person and income records in the SIPP relational microdata file, and at the end of appropriate sections within the records of the rectangular file, generally corresponding one-for-one with specific data items.

In the example in figure 4, the data item for earned income received from a particular job in a particular month is shown on the top half. A sample value of 2000 is illustrated, i.e., $2000 of income last month. Its corresponding imputation flag is shown on the bottom half. Note that the description of the imputation flag cites the field name for the corresponding item, WS1-2032. The sample value of 1 in the imputation flag indicates that the original respondent failed to answer the corresponding question, or the entry supplied was unusable for some reason, and that therefore the information in the data item above was imputed from that of another person.

In examining only the income amounts, one would not know that the $2000 was imputed rather than actually reported by the individual. Only by crosstabulating income by imputation status can one recognize an imputed income.

FIGURE 4. Illustration of an Imputation Flag

```
                    Data Dictionary                    Sample Values

                         (Wage and Salary Record)
Sample Data Item

        D WS1-2032         5     3293                     $2000
             What was the total amount of pay
             that ... received before deductions
             on this job last month (month 4).
             Range = -9,33332.
        U Persons 15 years old and older
        V -9.Not in universe
           0.None

Corresponding Imputation Flag

        D WS1CAL01         1     3321                         1
             Field 'WS1-2032' was imputed
        V 0 .No imputed input
          1 .Imputed input
```

Editing

There are also a number of demographic characteristics from the control card which should not require imputation, but may need to be edited for consistency with other information from the household. In these cases there are no imputation flags, but the file includes both the edited value and the value prior to computer editing, referred to as preedited or unedited. These items are identified by a "U" at the start of the 8-character mnemonic identifying variables in the data dictionary. To detect whether a particular edit had any impact on the data, compare a given data item with its preedited or unedited counterpart.

Uses of Imputation Flags

Although the Bureau could theoretically evaluate the above-cited sources of error--undercoverage, respondent and enumerator error, processing error and nonresponse--it does not do so for SIPP. Thus it is not possible to provide adjustment factors which could somehow be used to "correct" data. On the other hand, the user of the microdata files can study the impact of imputations made for nonresponse.

An analyst can use imputation flags or unedited items in several different ways. First, by computing the rate of imputation one can evaluate the quality of certain data items. For instance, one could find out whether persons receiving aid from the government are less likely to report their other sources of income than persons not participating in such programs.

Imputation flags allow characteristics of nonrespondents to be studied. Do nonrespondents tend to be younger or older, for example, than the rest of the population?

One can exclude imputed data from crosstabulations that might be sensitive to the imputation process. For instance, in comparing the earnings of doctors and dentists, high imputation rates might make the tabulations questionable, since missing income on a doctor's or dentist's record would be imputed from a pool of possible donors which includes a much broader range of professional occupations. Thus, to make sure you are comparing only doctor's incomes with dentist's incomes, it would be appropriate to exclude all cases with either occupation or income imputed.